Presentation
Agenda

High Productivity MPI – Grid, Multi-Cluster, and Embedded System Extensions

Pirabhu Raman, Anthony Skjellum, Rossen Dimitrov, Kumaran Rajaram, and ***Puri Banglore***
Verari Systems Software, Inc.
Phone: +1-205-314-3471
E-mail Addresses: {pirabhu, tony, rossen, kums, puri}@mpi-softtech.com

High Productivity MPI is an approach to extending MPI to support multiple

- Implementations (IMPI, IMPI-2)
- Owner domains
- Architectures
- Networks
- Operating Systems
- Faults
- Interacting dynamic groups

without relying on a two-level implementation (one MPI implementation calling another). MPI implementations must be able to connect, reconnect, and work well with dynamic, intermittent resources, under the expectation that user applications will also become somewhat fault-aware in order to retain scalability.

This paper addresses the many concerns that arise in offering composable sessions in which multiple-vendor MPI's can be supported (starting from but not ending with IMPI protocol). Experiences with IMPI, and a new proposal, IMPI-2, are offered. This paper addresses specific issues about interoperating the gamut of MPI-2 services in the interoperable setting, which to our knowledge have not been addressed elsewhere.

The results of this work are open specifications, together with our own vendor-implementation of these MPI capabilities. Other open and commercial MPI's could adopt IMPI plus these other extensions in order to participate in the hierarchical, heterogeneous, grid computing settings, without mandating new MPI implementations in such settings. The authors' goals in offering these new protocols as proposals is to encourage the High Productivity Computer world to enter into significant discussions about their adoption. The goal is to offer these capabilities without mandating grid-computing infrastructure.

Specific requirements for supporting several overlapping and non-overlapping networks of varying performance (overhead, bandwidth, latency, concurrency) are discussed, in terms of progressive MPI implementations, and the joint progressive nature of compliant MPI's working together with the IMPI-2 protocol framework. Note that existing multi-cluster/grid solutions apparently cause excessive polling, and do not support the degree of scalability or appropriate intra-network performance that would otherwise pertain to correctly composed MPI implementations.

In the spirit of pursuing practical fault tolerance in this setting, the extension of checkpoint-restart facilities to High Productivity MPI is considered both from the perspective of MPI I/O on

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **01 FEB 2005**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
High Productivity MPI Grid, Multi-Cluster, and Embedded System Extensions

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Verari Systems Software, Inc.

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001742, HPEC-7 Volume 1, Proceedings of the Eighth Annual High Performance Embedded Computing (HPEC) Workshops, 28-30 September 2004. , The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**17**	

single implementations, and in terms of the MPI I/O as extended to multiple implementations connected through an IMPI-2 protocol framework.

Other aspects of fault-tolerant MPI are beyond the scope of this paper.

The authors note that there are many thorny issues associated with supporting all the features of MPI-2 across multiple vendor implementations, and several of these issues are highlighted. A "core MPI-2" (subset of MPI-1 plus subset of MPI-2) plus some additional dynamic processing extensions are suggested as a best practice for computing in this setting.

Networks of grids (or multi-clustering) represent an interesting capability but also a challenge in terms of the publication of the entire structure (including IP addresses) globally. This work also considers the use of techniques such as NAT and port forwarding, together with gateway nodes, in order to allow for structured, manageable descriptions of hierarchical parallel resources, so that appropriate communication bandwidth remains possible between clusters, without mandating a public model for all resources involved. This has been accomplished with IMPI as is, and extensions in the IMPI-2 framework are discussed.

The paper also tries to address some of the drawbacks with existing IMPI protocols including

- Involvement of user in starting MPI jobs distributed across multiple platforms
- Global collective operations mandated by standard
- Non-portable parallel job startup mechanism

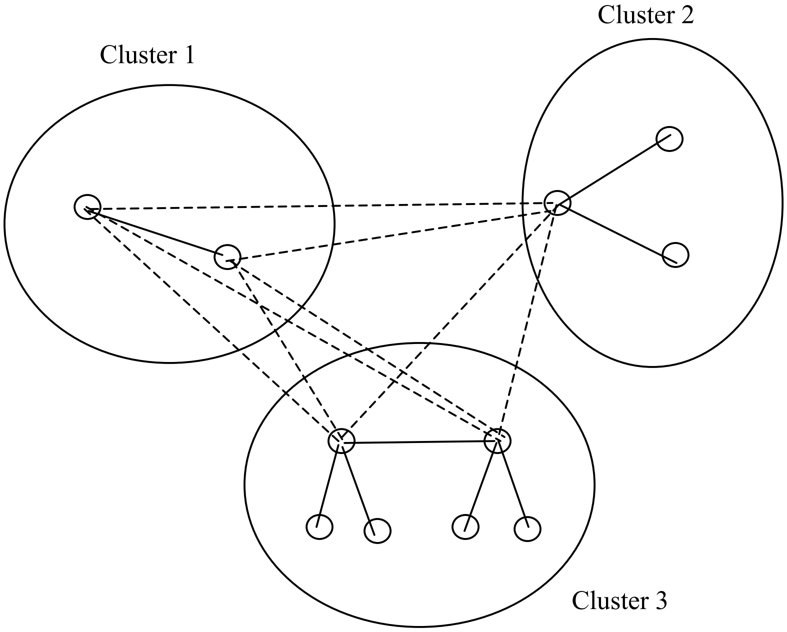


Figure 1: Connectivity Architecture of a multi-cluster, multi-implementation MPI

Figure1 shows a typical scenario where multiple-clusters (possibly under different owner domains and different internal networks) having multiple implementations of MPI. Our approach enables these multiple implementations to work seamlessly without requiring a new-layered implementations and also enables individual implementations to use proprietary and optimized communication protocols with no increased overhead. The IMPI-1 implementation, which is currently available as part of MPIPro, was supported by NIST through the Contract # 50-DKNB-1-SB082.

References:

1. MPI Forum. *MPI: A Message-Passing Interface Standard*. 1994. http://www.mpi-forum.org/docs.
2. IMPI Steering Committee. *IMPI Draft Standard*. 1999. http://www.nist.gov/impi/.
3. Gropp, W. et al. A High-Performance Portable Implementation of the Message-Passing Interface. *Journal of Parallel Computing*, 22, 1996, pp: 789-828.
4. Skjellum, A. and McMahon, T. *Interoperability of Message-Passing Interface Implementations: a Position Paper.* http://www.mpi-softtech.com/company/publications/files/interop121897.pdf

Abstract

Agenda

High Productivity MPI – Grid, Multi-cluster and Embedded Systems Extensions

Presented by

Dr. Anthony Skjellum

Chief Software Architect

Verari systems Software, Inc

September 30, 2004

HPEC Workshop

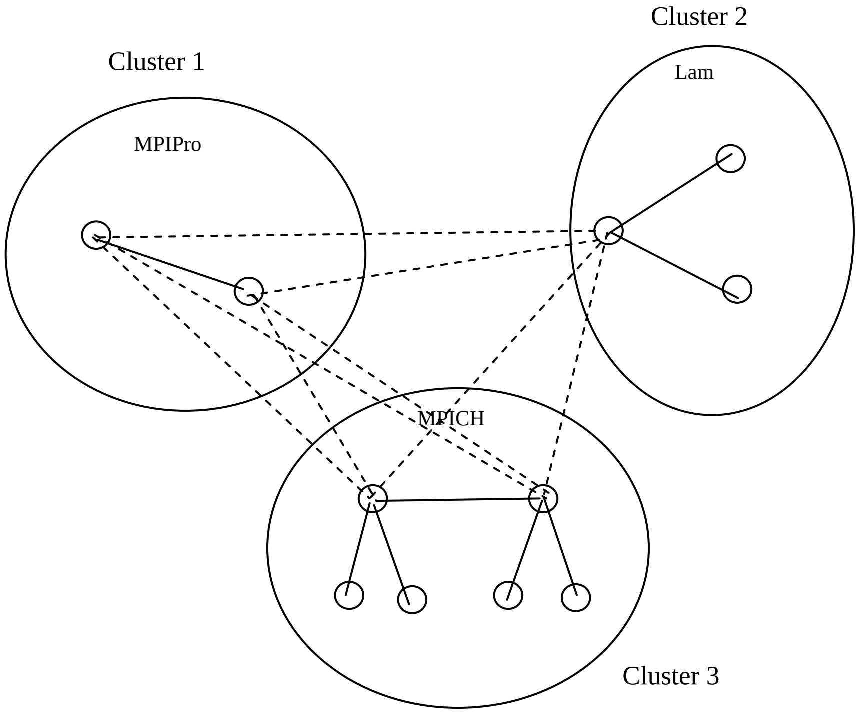
IMPI

- **IMPI – Interoperable Message Passing Interface**
- **Developed and Proposed by NIST**
- **Standard for inter-operation of multiple**
  - **Implementations (IMPI, IMPI-2)**
  - **Architectures**
  - **Networks**
  - **Operating Systems**

Client, Host Concept

- **MPI processes spread across multiple clients**
- **Clients represent MPI processes belonging to a single implementation**
- **Hosts represent gateways for processes of Clients**
- **IMPI Application may have two or more clients**
- **Client may have one or more hosts**
- **Hosts serve as gateways for one or more MPI processes**

Typical Scenario – Multi-vendor MPI



- 3 Clients (Each cluster make one client)
- Client 1
    - 2 hosts, 2 MPI processes
- Client 2
    - 1 host, 3 MPI processes
- Client 3
    - 2 host, 6 MPI processes

MPI/Pro 1.7.0

- MPI/Pro 1.7.0 provides first complete implementation of IMPI
- Enables Interoperation between
  - Windows, Linux and Mac OSX operating systems
  - 32-bit and 64-bit architectures
  - TCP, GM and VAPI Networks
  - Any combination of all the above

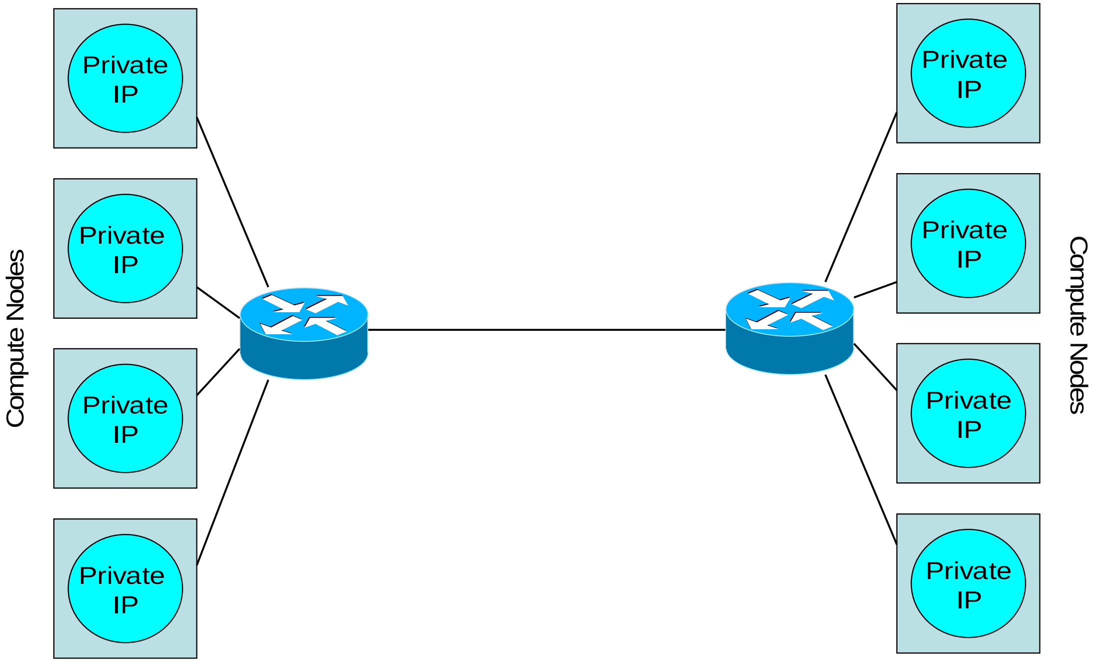
Extensions

- IMPI does not address issues such as
  - Private IP Addresses
  - Owner domains
  - Faults
  - Interacting Dynamic Groups
- Above issues play vital role in Grid
- Verari proposed and implemented a novel method to address issue of Private IP Addresses

Case Study

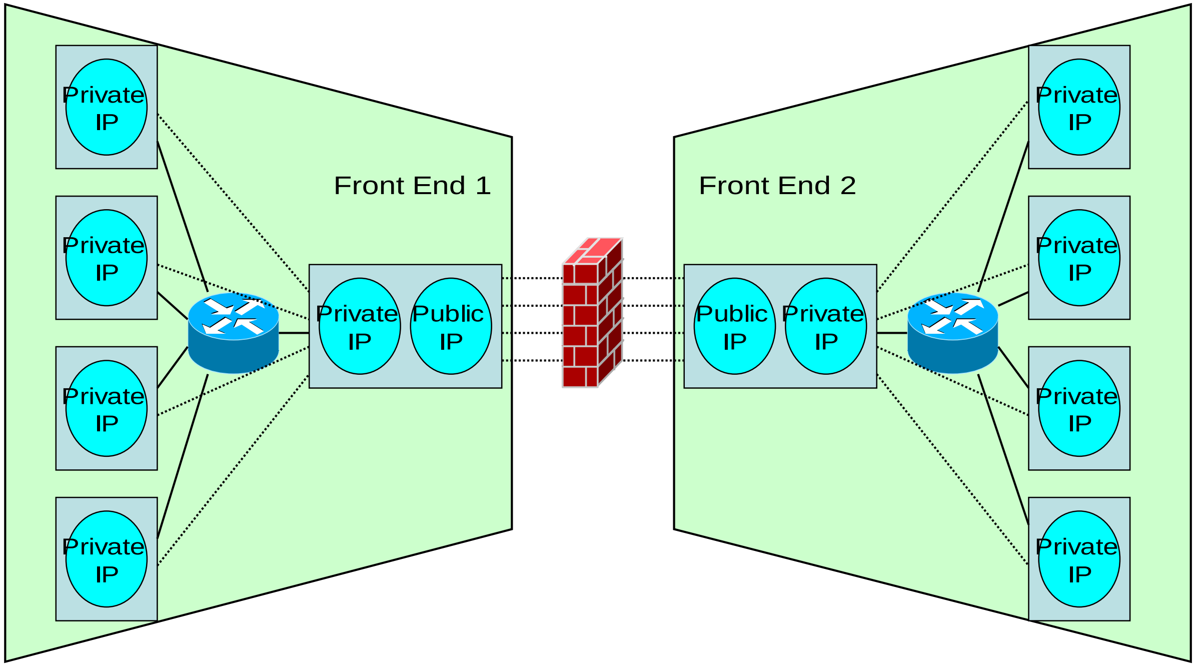
Private IP Enabled IMPI

Typical Cluster Setup



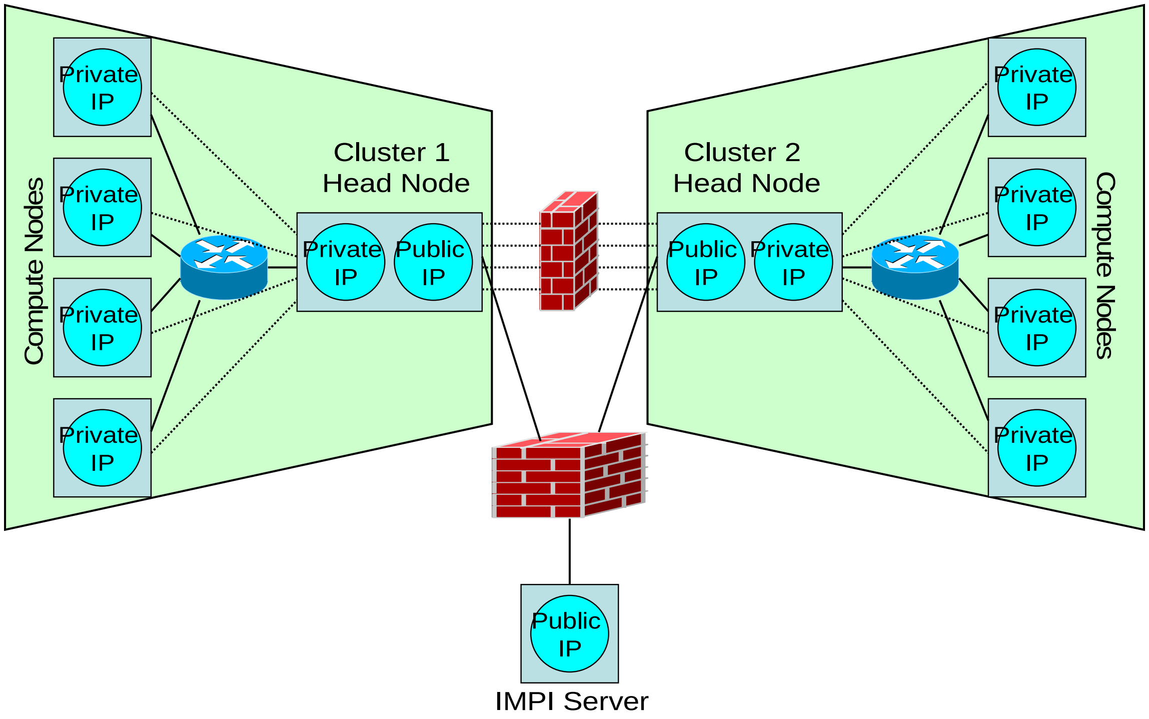
- Compute Nodes have private IP addresses
- External communication through single head node or gateway
- Unsuitable for multi-cluster multi-site communication

Network Address Translation (NAT)



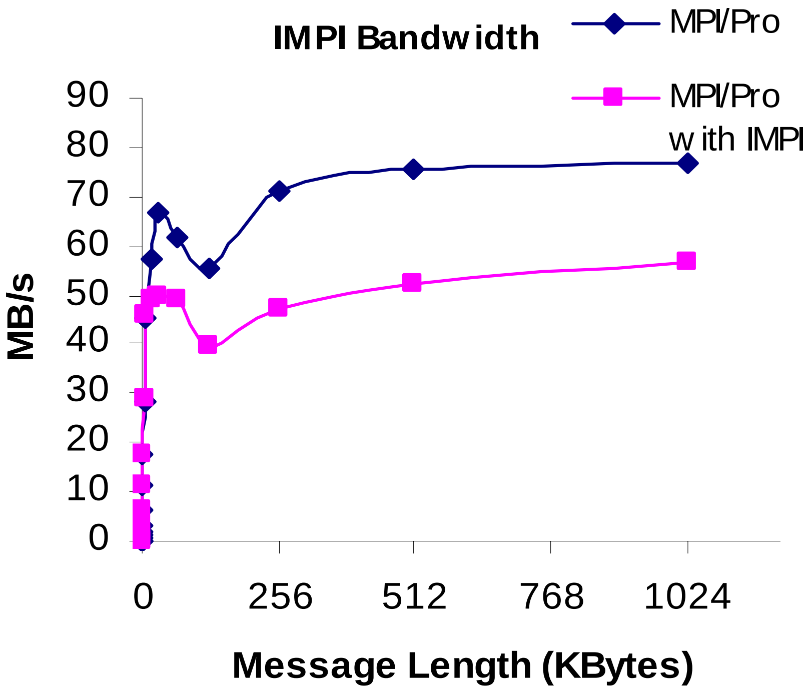
- Firewalls block incoming connections
- NAT used to serve requests generated within the cluster

NAT-based IMPI



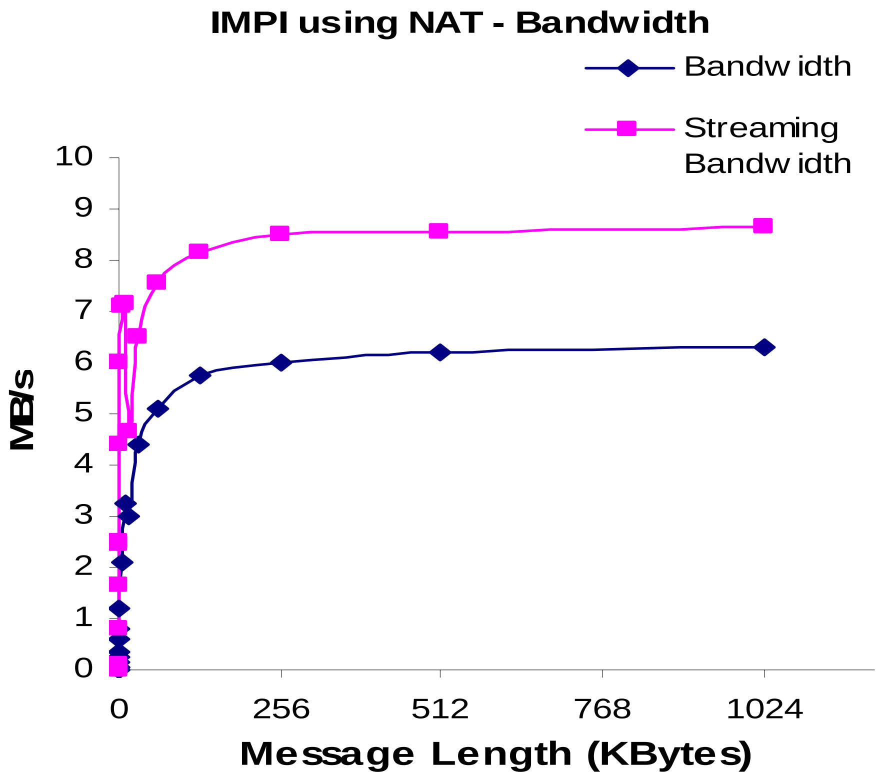
- Use NAT to generate dynamic mappings between head node and compute nodes
- Dissipate dynamic mapping info through IMPI server
- Release mapped ports on head node on completion of application

Performance



Configuration	Latency (us)
MPI/Pro without IMPI	142.45
MPI/Pro with IMPI	147.35

Performance



10/13/2004

Proposed Extensions

- IMPI extensions for MPI-2
- Open protocol-based initialization such as SOAP
- Adaptation to the Grid
- Reduce user involvement
- Optimize for performance

References

- Velusamy, V *et al. Communication Strategies for Private-IP-Enabled Interoperable Message Passing across Grid Environments*, First International Workshop on Networks for Grid Applications, 2004.
- MPI Forum. *MPI: A Message-Passing Interface Standard*. 1994. http://www.mpi-forum.org/docs.
- IMPI Steering Committee. *IMPI Draft Standard*. 1999. http://impi.nist.gov/IMPI/.
- *Gropp, W. et al. A High-Performance Portable Implementation of the Message-Passing Interface.* Journal of Parallel Computing, 22, 1996, pp: 789-828.
- Skjellum, A. and McMahon, T. *Interoperability of Message-Passing Interface Implementations: a Position Paper.* http://www.verarisoft.com/publications/files/interop121897.pdf